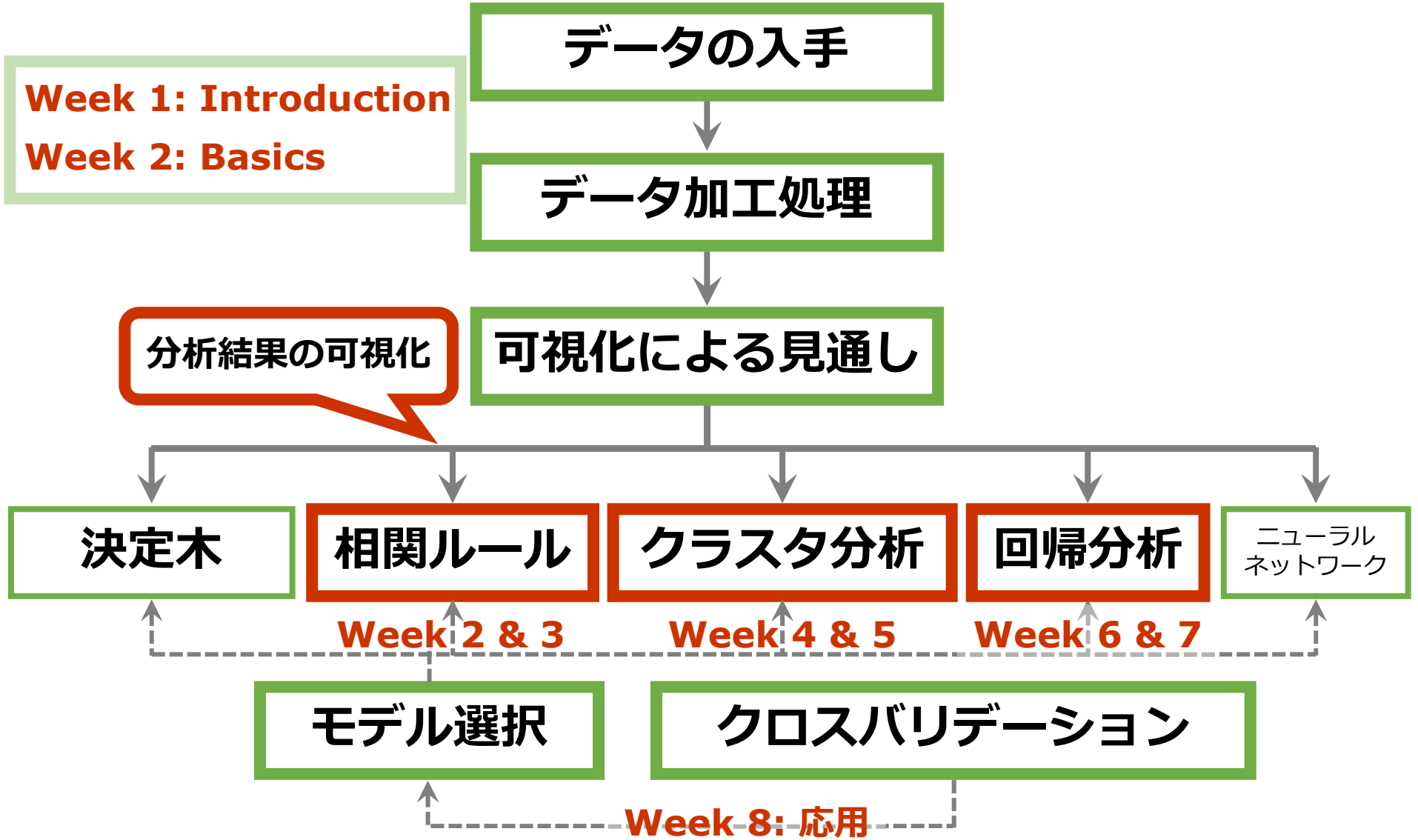
データマイニングと情報可視化

Week 8

稲垣 紫緒
いながき　しお

理学研究院　物理学部門 / 共創学部
inagaki@phys.kyushu-u.ac.jp
ウェスト1号館 W1-A823号室

授業計画


Week 1: Introduction
Week 2: Basics
データの入手
データ加工処理
分析結果の可視化
可視化による見通し
決定木
相関ルール
クラスタ分析
回帰分析
ニューラル
ネットワーク
Week 2 & 3
Week 4 & 5
Week 6 & 7
モデル選択
クロスバリデーション
Week 8: 応用

データマイニングの代表的な手法

(1) マーケットバスケット分析

どの商品とどの商品を
どのような顧客が同時に購入したかを分析

店内の陳列方法を改善

amazon

楽天 ICHIBA

この商品をを買っている人は
これも買ってます

マーケットバスケット分析

(1) 支持度

support

$$\frac{A \cap B}{N}$$

(2) 確信度

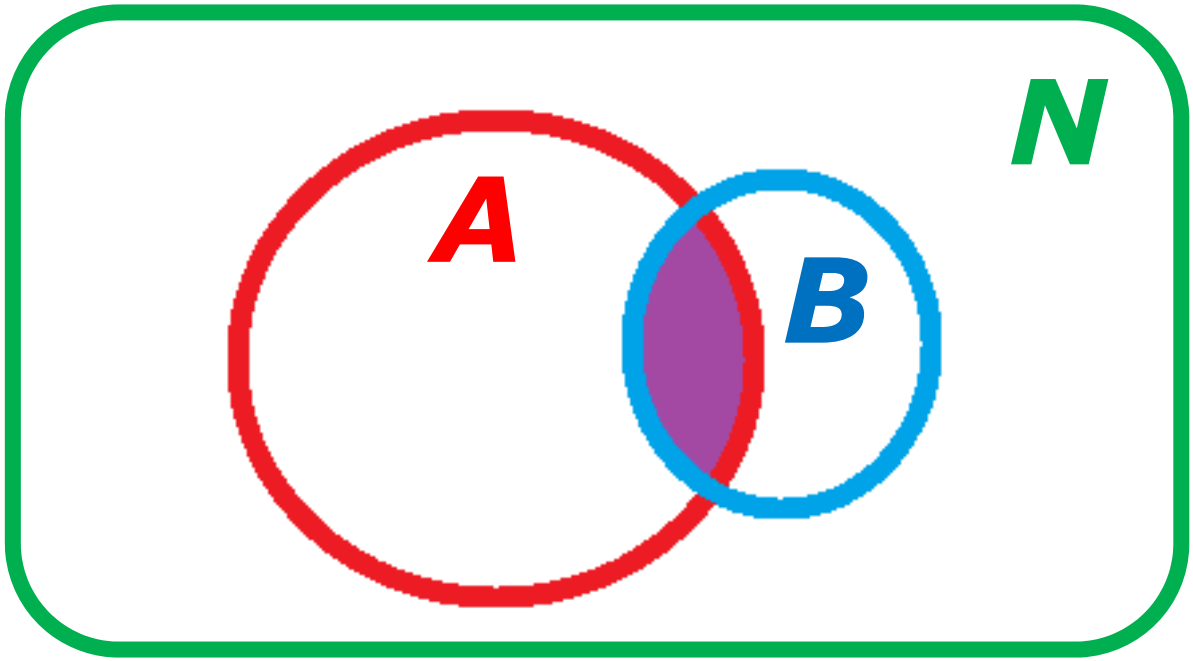
confidence

$$\frac{A \cap B}{A}$$

(3) リフト値

lift

$$\frac{A \cap B}{A\,B/N}$$



マーケットバスケット分析

(1) 支持度

support

$$\frac{A \cap B}{N}$$

(2) 確信度

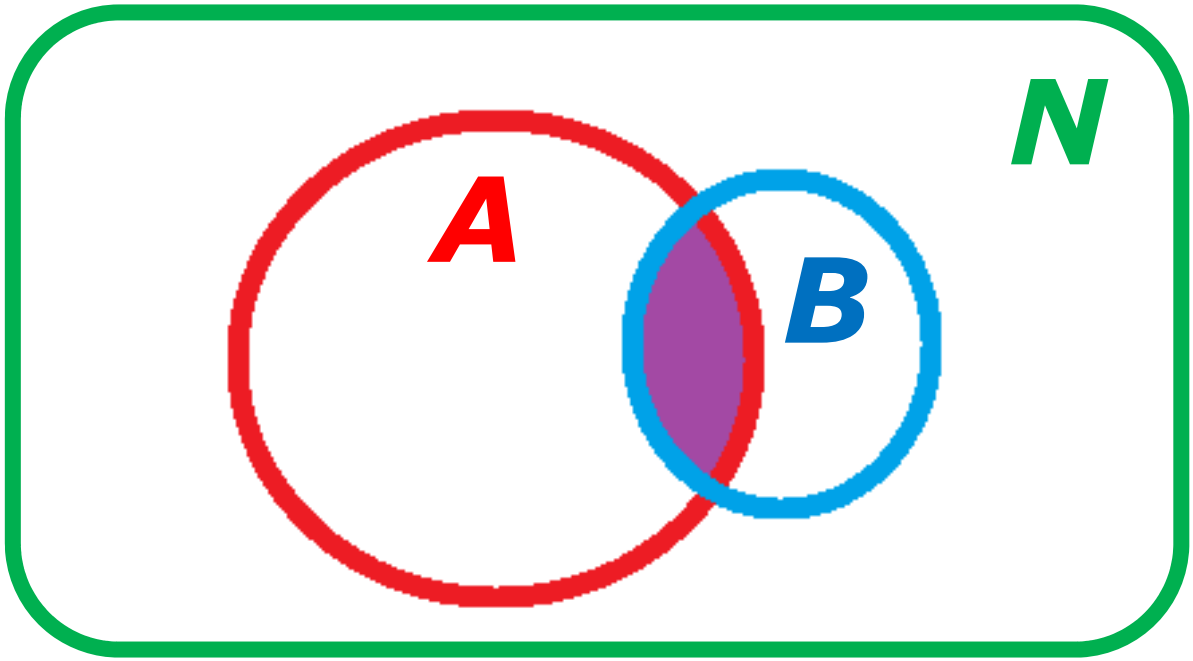
confidence

$$\frac{A \cap B}{A}$$

(3) リフト値

lift

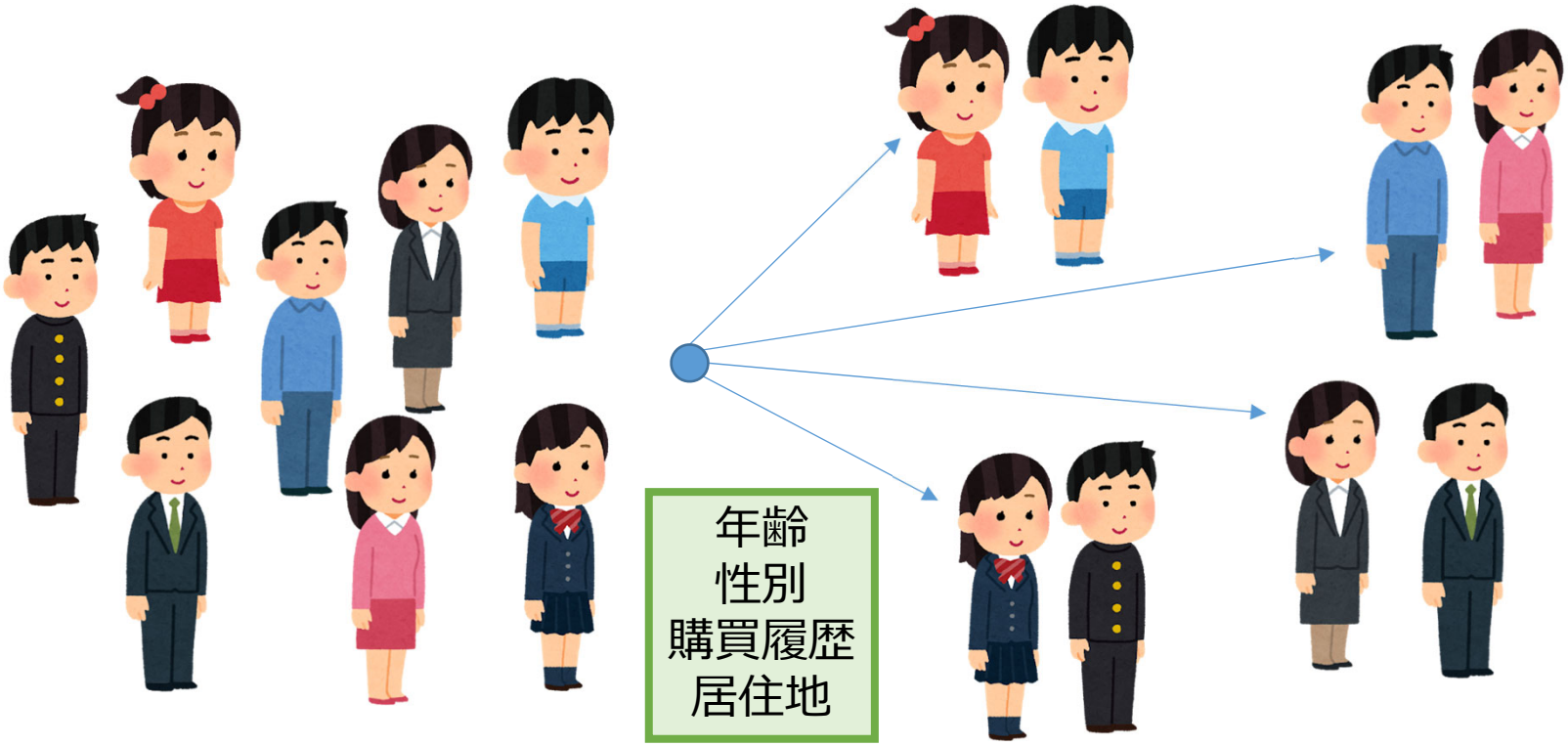
$$\frac{A \cap B}{A\,B/N}$$



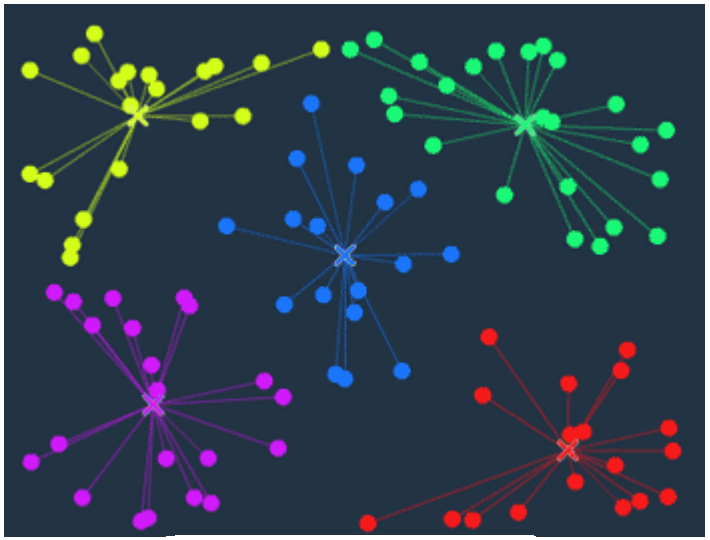
データマイニングの代表的な手法

(2) クラスター分析

似ているデータごとにデータをまとめて分類
→適切な商品を推奨できる



k-means法



nstep=9

クラスタ数=5

step 0 クラスタの数を決める

step 1
各点にランダムにクラスタを割り当てる

step 2
クラスタの重心を計算

変化あり
2 に戻る

step 3
点のクラスタを、
一番近い重心のクラスタに変更する

クラスタの
組み換えなし

終了

データマイニングの代表的な手法

(3) ロジステック回帰分析

発生確率を予測する

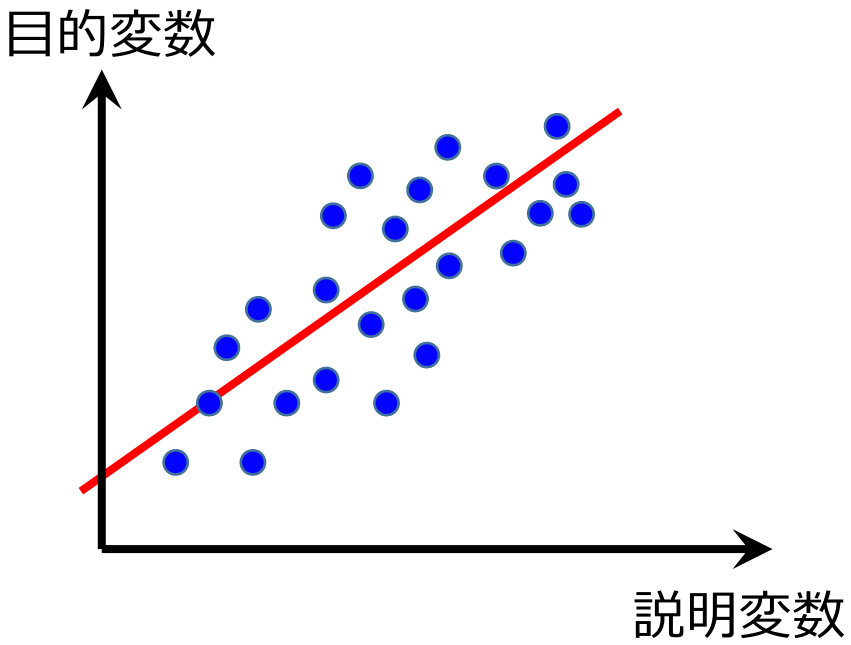
■ がんの発症確率や生存率など
■ アンケート結果から、携帯会社を乗り換える顧客を予測

au / AEON MOBILE

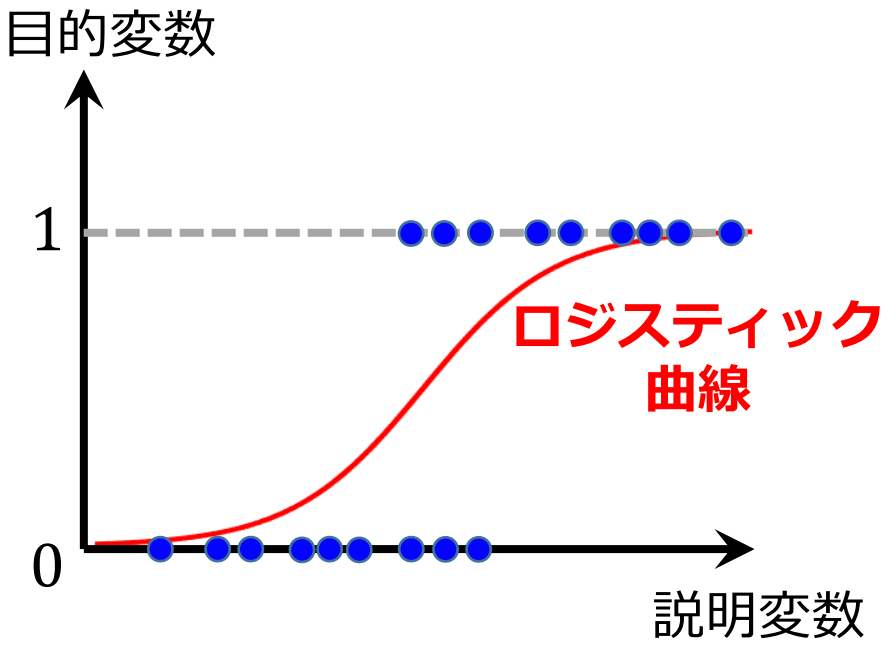
NTT docomo = SoftBank

回帰分析

線形回帰分析
→**量的変数**の予測



ロジスティック回帰分析
→**発生確率**の予測



ある事象が　**起きた→ 1**
起きなかった→0

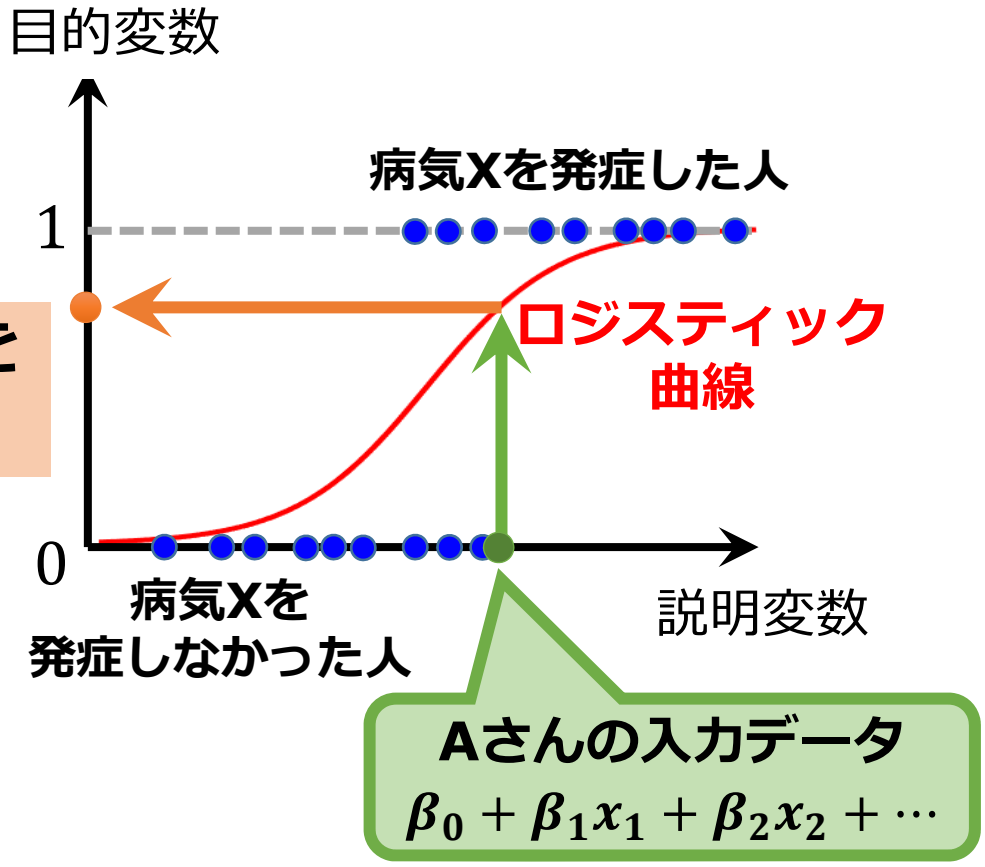
ロジスティック回帰分析

ある事象が　**起きた→ 1**
起きなかった→0

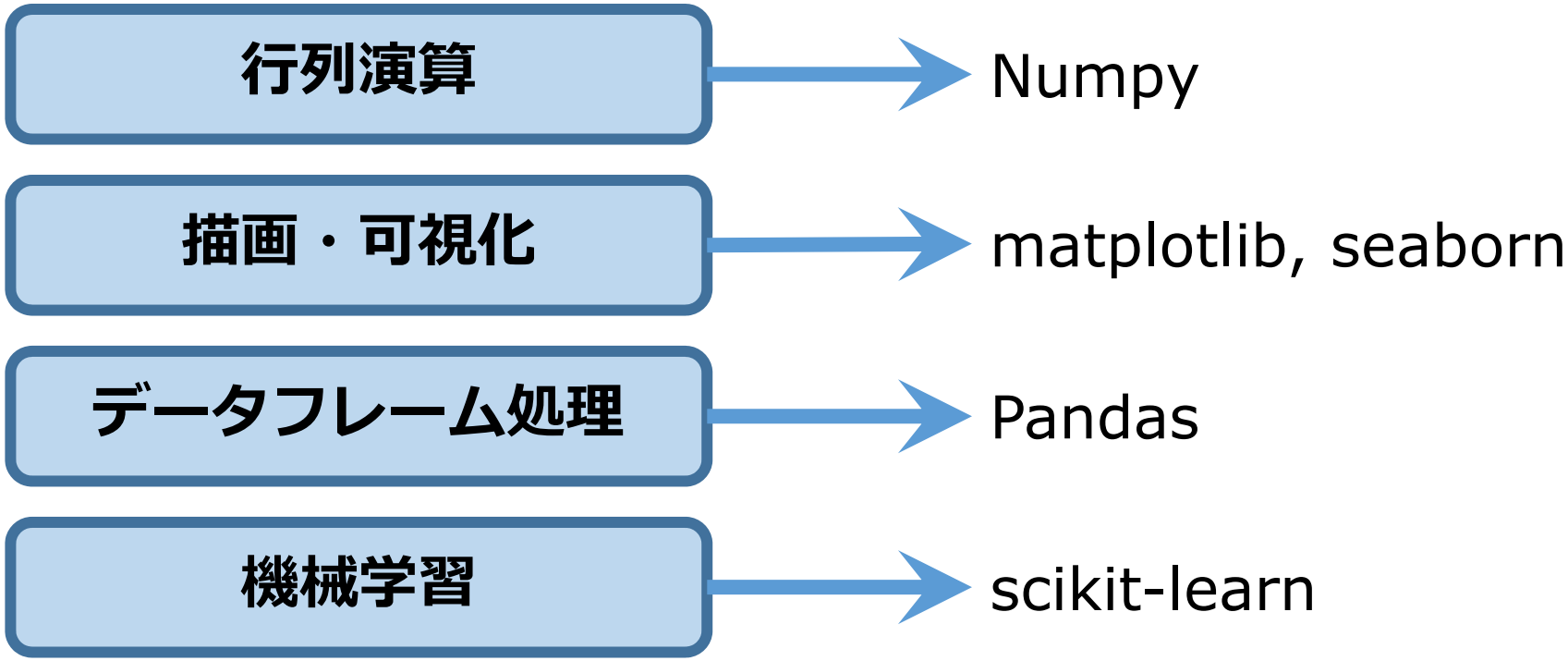
説明変数
- アルコール摂取量
- 喫煙歴
- 体脂肪率
- BMI
- 年齢

Aさんが病気Xを発症する確率

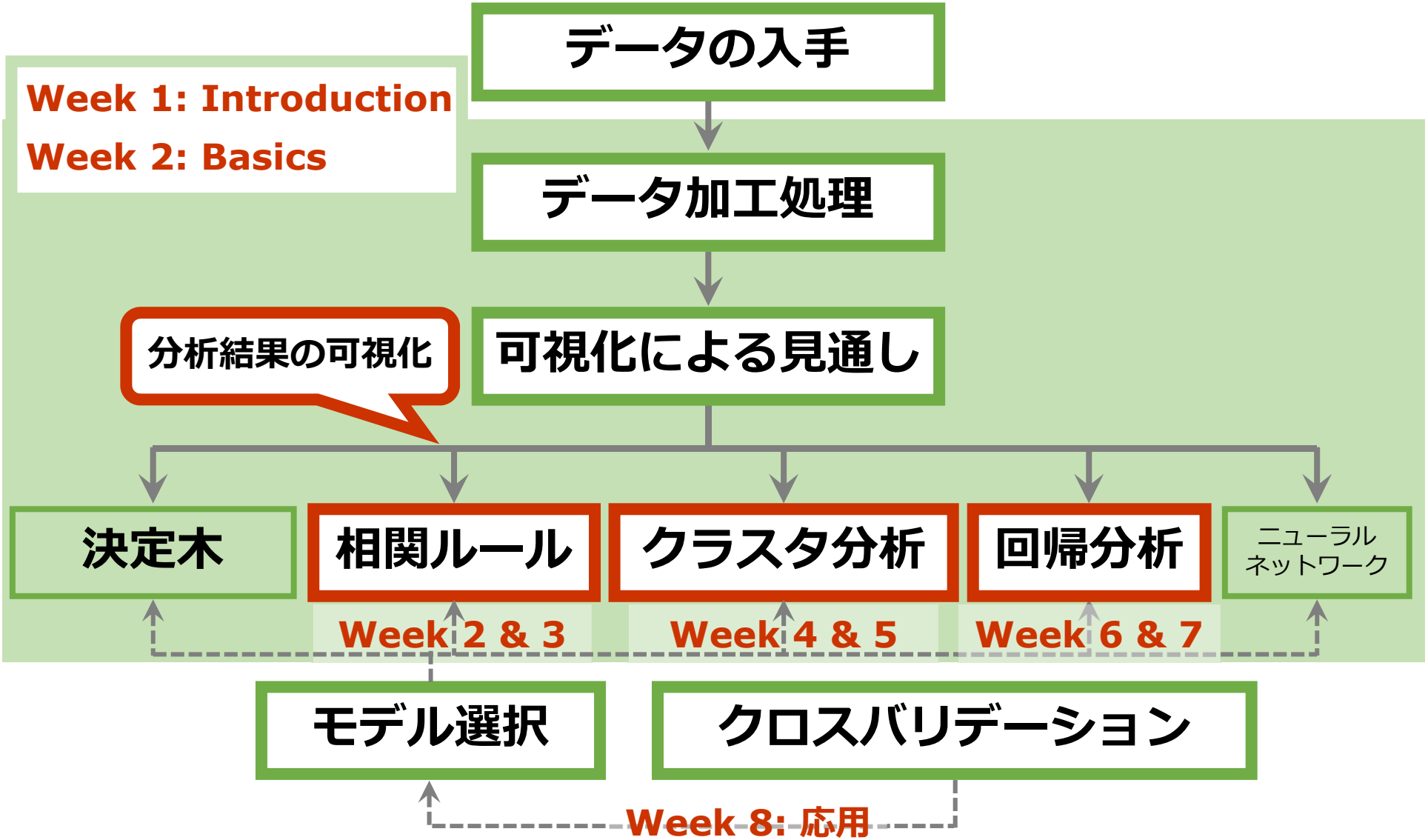
ロジスティック回帰分析
→**発生確率**の予測



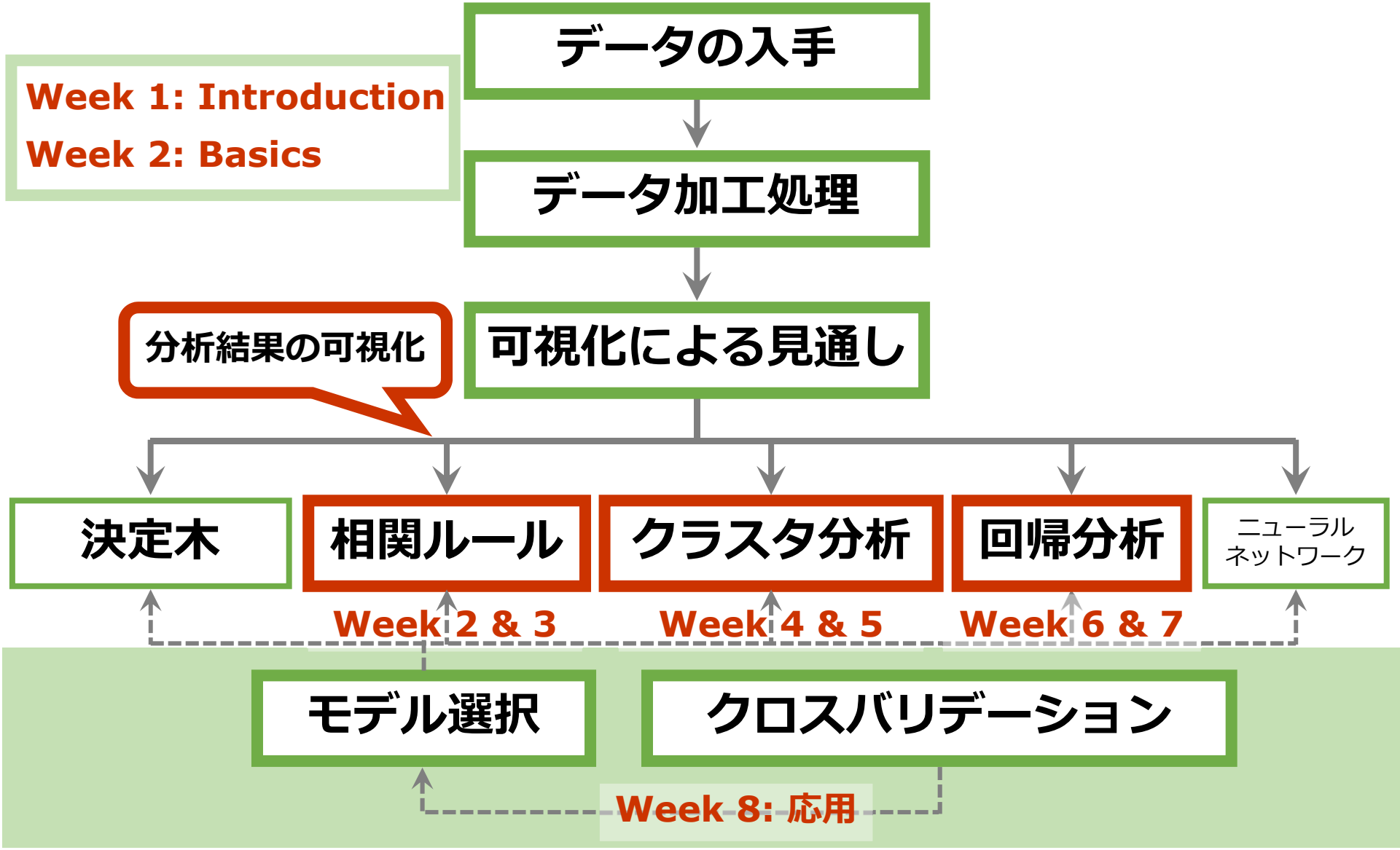
Pythonのライブラリ



授業計画

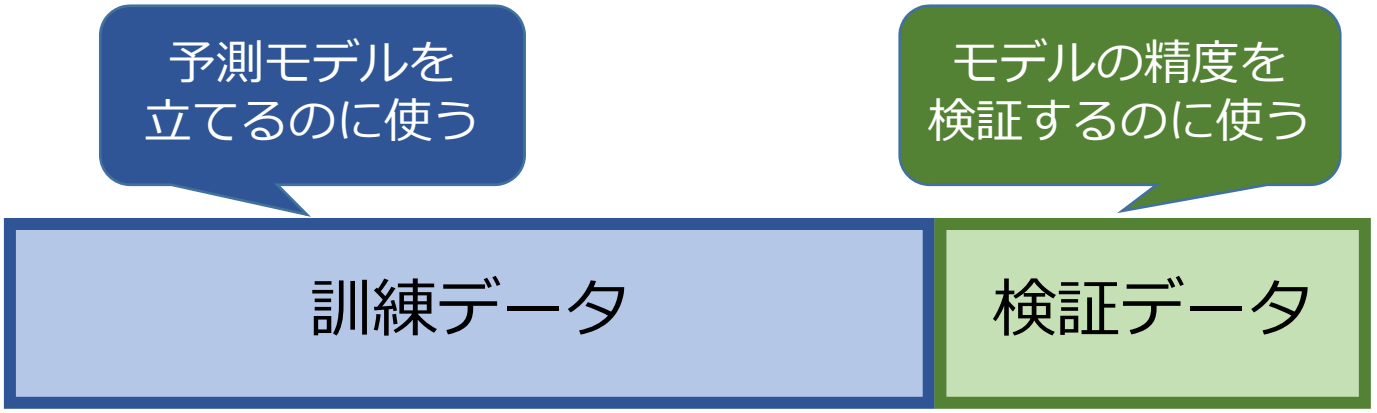

データの入手
Week 1: Introduction
Week 2: Basics
データ加工処理
分析結果の可視化
可視化による見通し
決定木
相関ルール
クラスタ分析
回帰分析
ニューラル
ネットワーク
Week 2 & 3
Week 4 & 5
Week 6 & 7
モデル選択
クロスバリデーション
Week 8: 応用

授業計画


Week 1: Introduction
Week 2: Basics
データの入手
データ加工処理
分析結果の可視化
可視化による見通し
決定木
相関ルール
クラスタ分析
回帰分析
ニューラル
ネットワーク
Week 2 & 3
Week 4 & 5
Week 6 & 7
モデル選択
クロスバリデーション
Week 8: 応用

汎化性能（未知のデータに対する性能）

ホールドアウト法

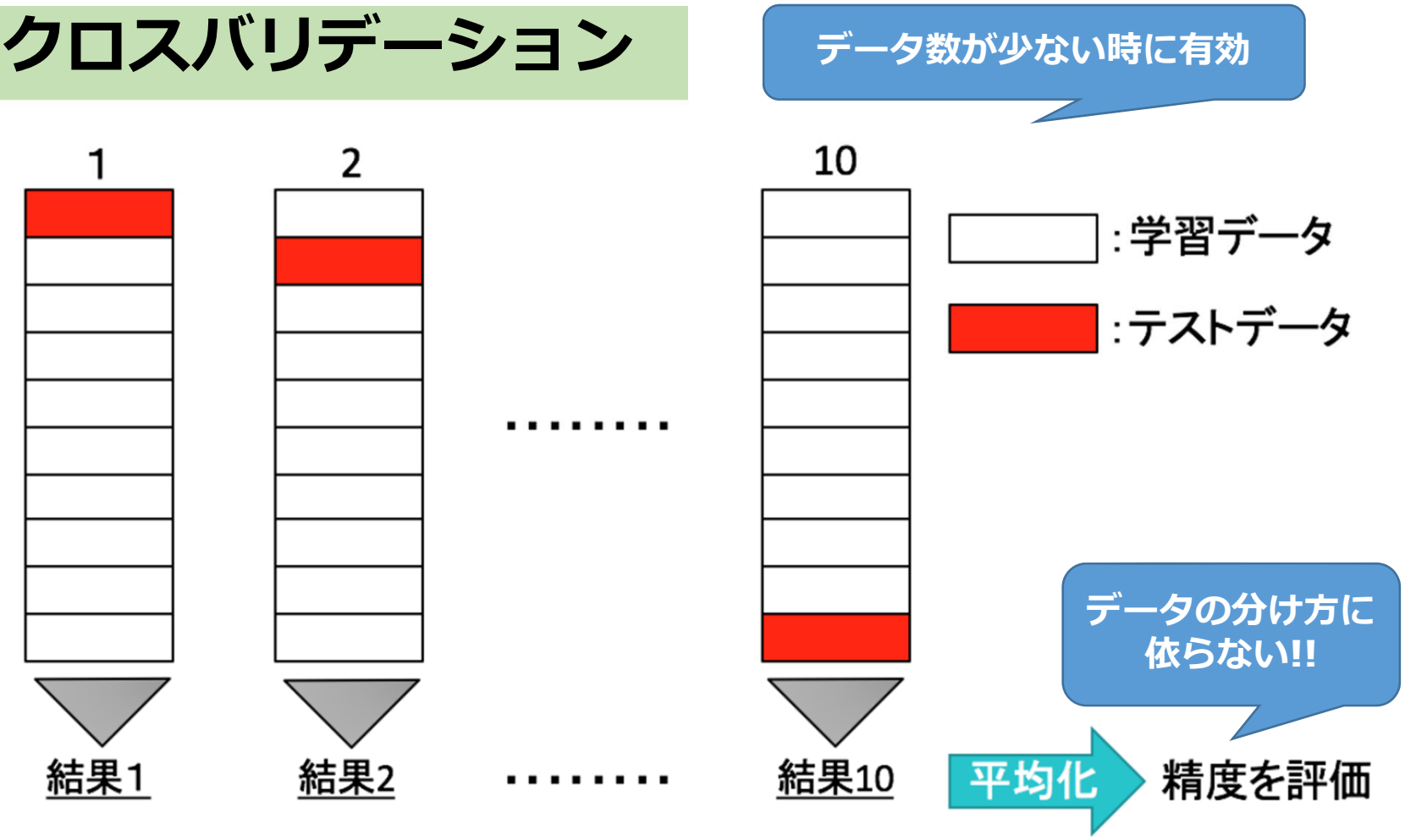


データの分け方によって正解率が変わる。
random_state で、データの分け方を変えられる。

データの数が十分に大きければ、正解率は大きく変わらない。

汎化性能（未知のデータに対する性能）

クロスバリデーション



https://aiacademy.jp/texts/show/?id=299

モデル選択

いくつかのモデルで比較して、
精度の良いモデルを採用する。

分類→
- k-means法
- 混合ガウスモデル
- 決定木
- サポートベクターマシン

正解率

モデルの性能評価

$$\textbf{正解率(Accuracy)}=\frac{TP+TN}{Total}$$

検出率、**精度**など
他にもあります。

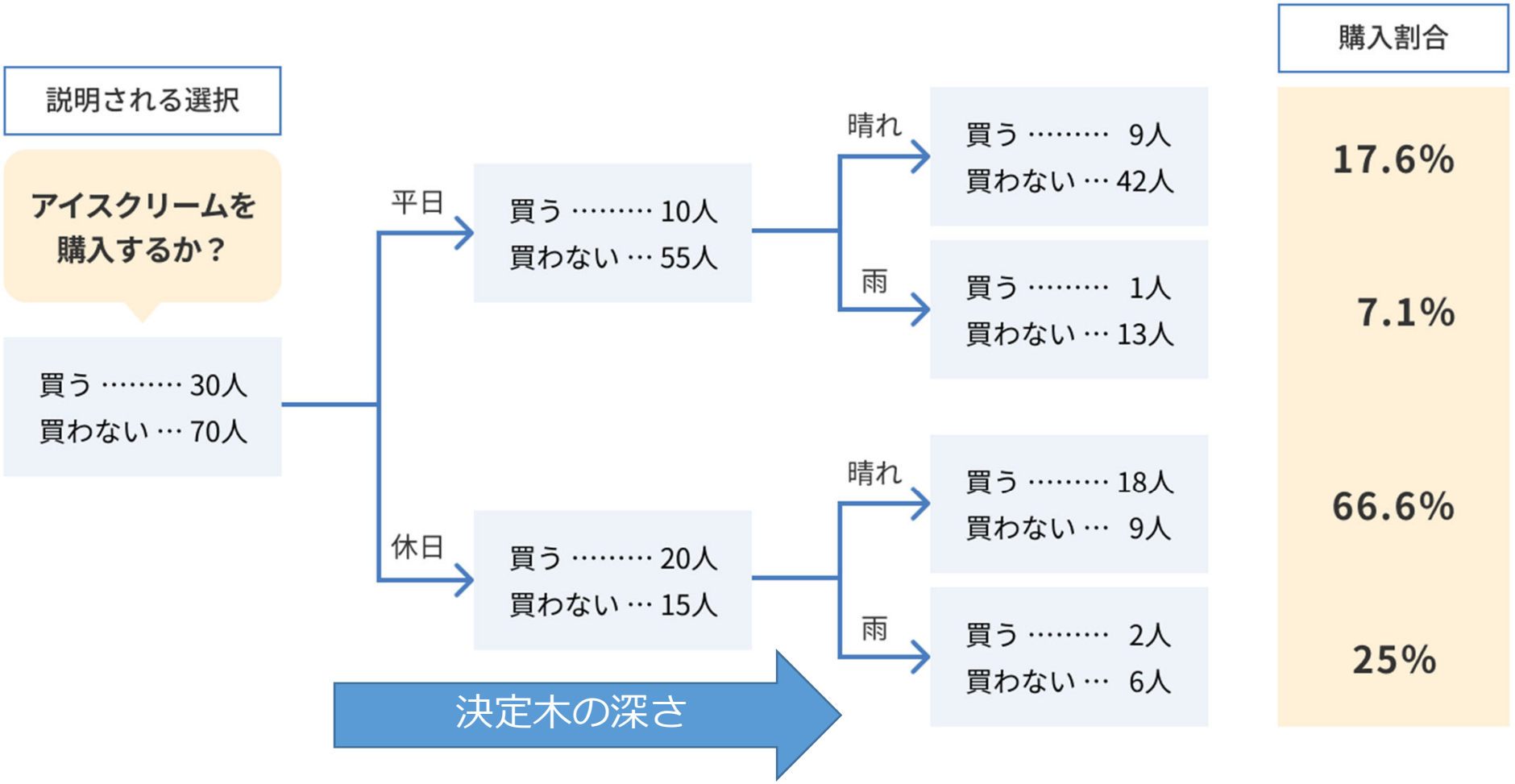
いろいろありますが、とりあえず正解率で評価

```
from sklearn import metrics
metrics.accuracy_score(iris_ans.species, df_pred)
```

正解　　予測

決定木（分類木）

アイスクリームを買うか？買わないか？



https://cacoo.com/ja/blog/what-is-decision-tree/

決定木で分類：アヤメのデータ

深さの決め方は
今回は割愛します。。

k-means法と
正解率を
比べてみましょう!!

Scikit-learnで決定木

```
from sklearn import tree
```

#インスタンスを作成

```
clf=tree.DecisionTreeClassifier(max_depth=3)
```

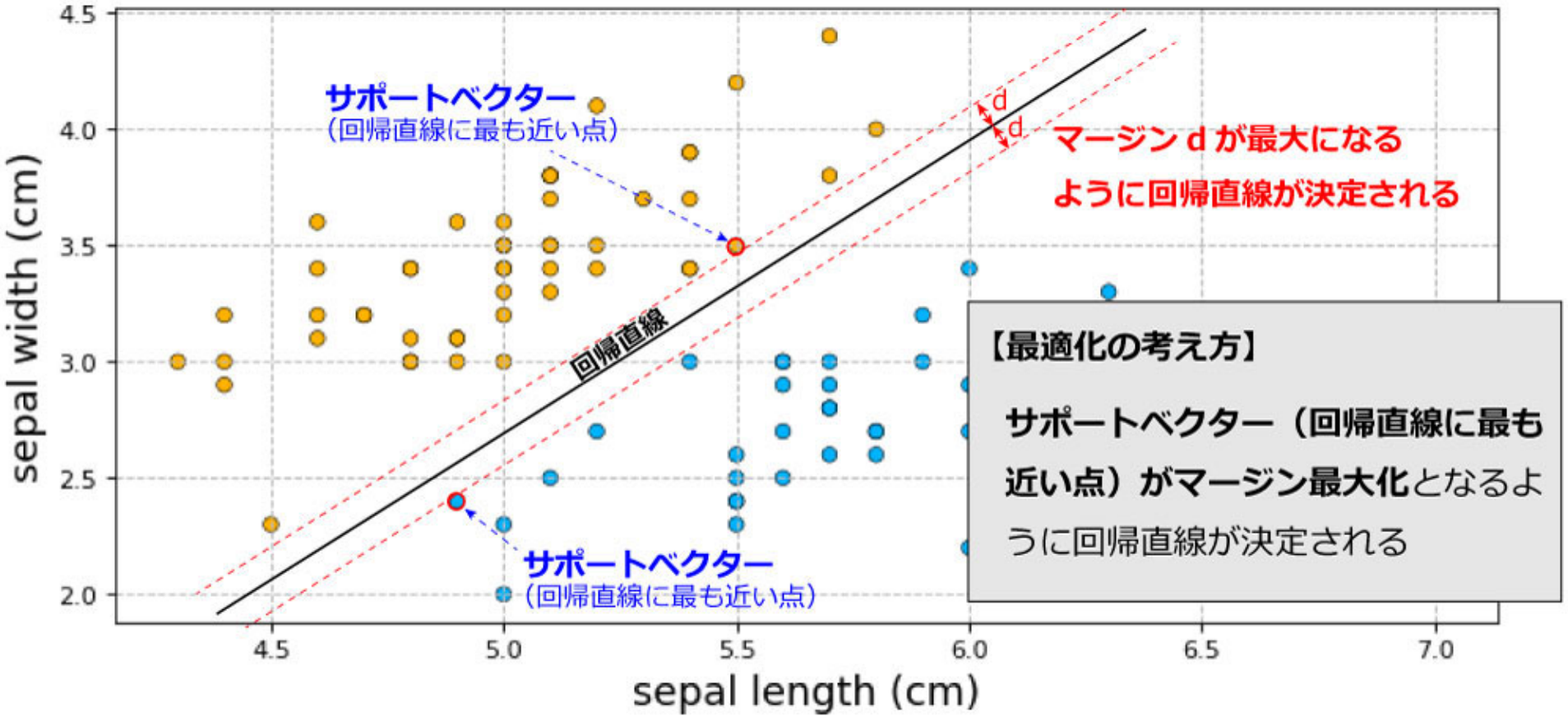
初期値はC=1.0 # データ学習

```
clf = clf.fit(X_train, y_train)
```

学習したモデルを使いテストデータで予測

```
clf.predict(X_test)
```

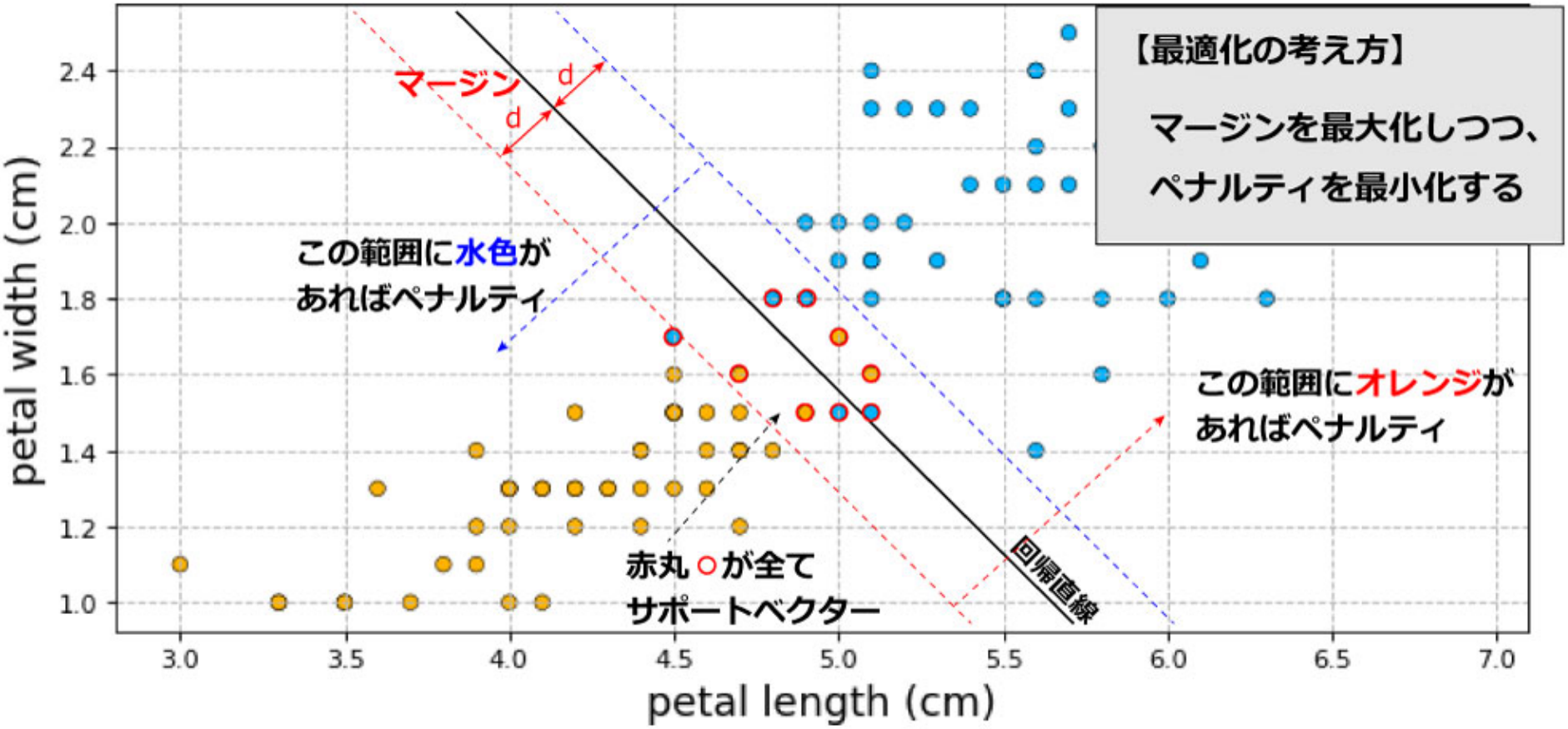
サポートベクターマシン(SVM)



基本的に２クラス分類に特化

https://kit.socinno.com/2_3_ai/

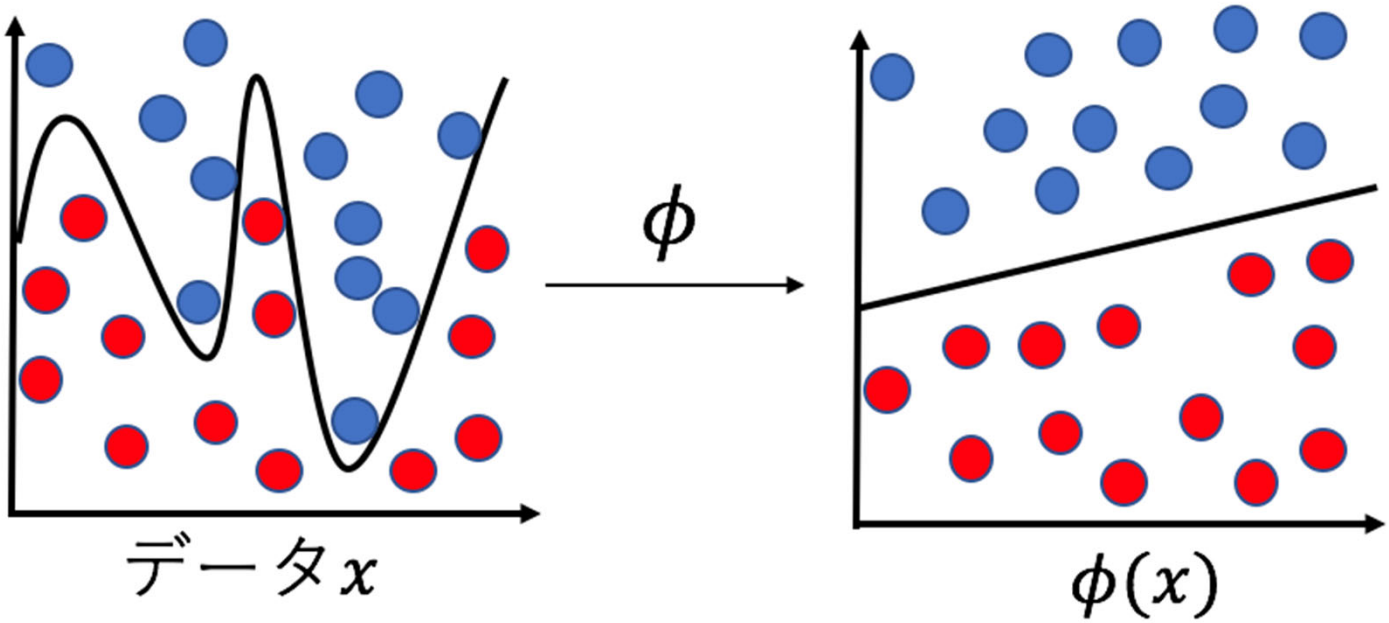
サポートベクターマシン(SVM)



https://kit.socinno.com/2_3_ai/

サポートベクターマシン(SVM)

- 基本的に２クラス分類に特化
- 多変数の分類が得意
- 一見複雑に入り組んでるデータでも、非線形カーネルを使ってうまく分類できる
- データ量が多いと計算量が多くて大変



Scikit-learnでSVM

```
from sklearn import svm,metrics
```

#インスタンスを作成

```
clf = svm.LinearSVC()
```

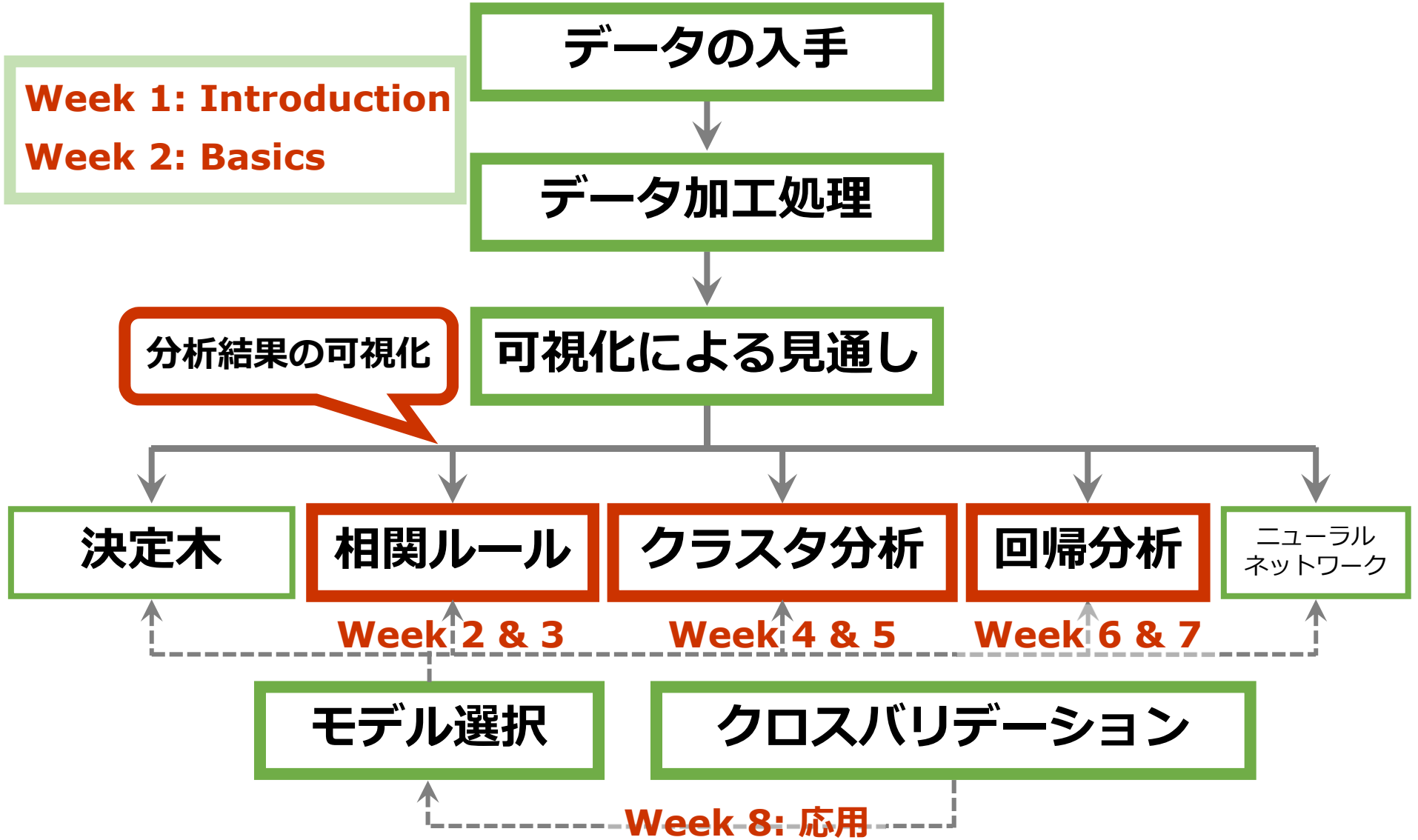
初期値はC=1.0 # データ学習

```
clf.fit(X_train, y_train)
```

学習したモデルを使いテストデータで予測

```
clf.predict(X_test)
```

授業計画


Week 1: Introduction
Week 2: Basics
データの入手
データ加工処理
分析結果の可視化
可視化による見通し
決定木
相関ルール
クラスタ分析
回帰分析
ニューラル
ネットワーク
Week 2 & 3
Week 4 & 5
Week 6 & 7
モデル選択
クロスバリデーション
Week 8: 応用

リフレクションシート

締め切りは本日23:59です。

授業アンケートも答えてください。

小テストの締め切りは
10:10
です。